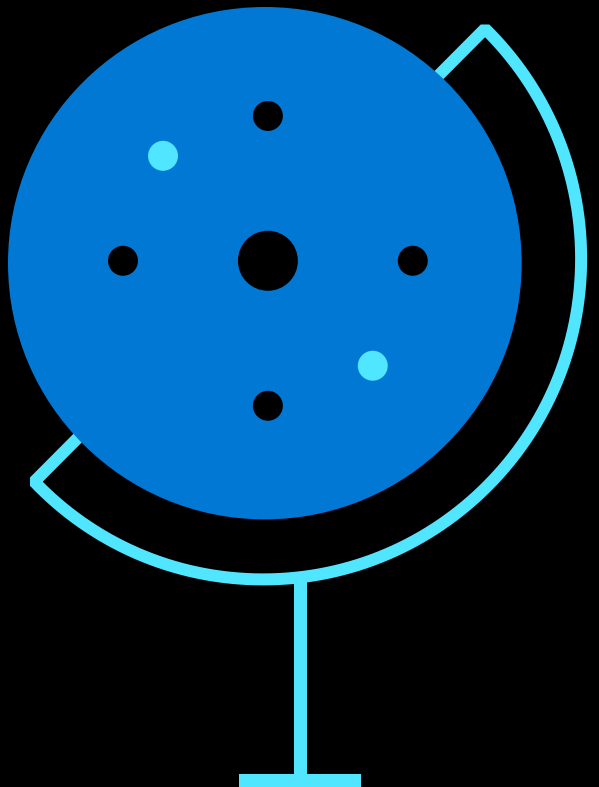
Azure and Alveo: Cloud Migration Made Easy

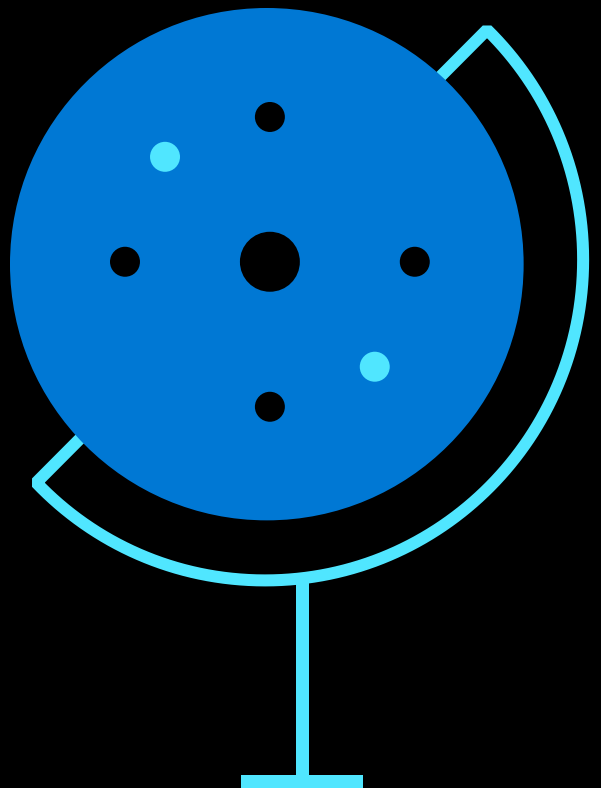
...with the NP-series VM platform.

Xilinx Adapt 2021

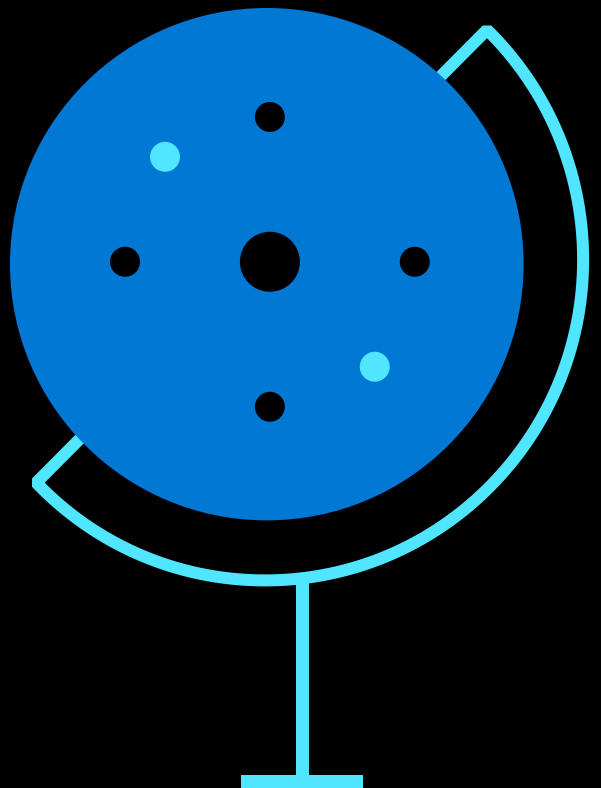
?

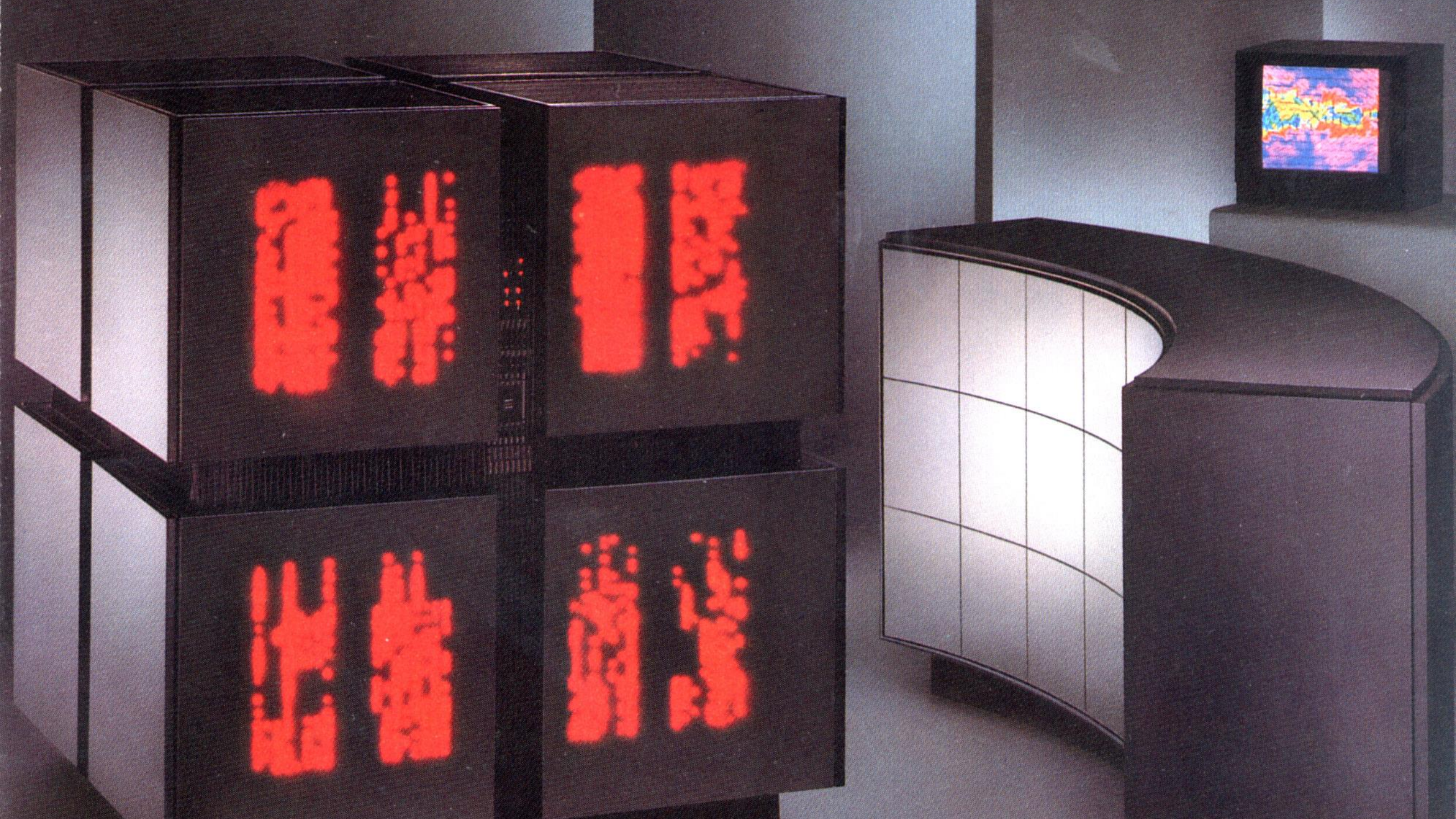


2.5 Exabytes (quintillion bytes) of new data is produced and stored every day.

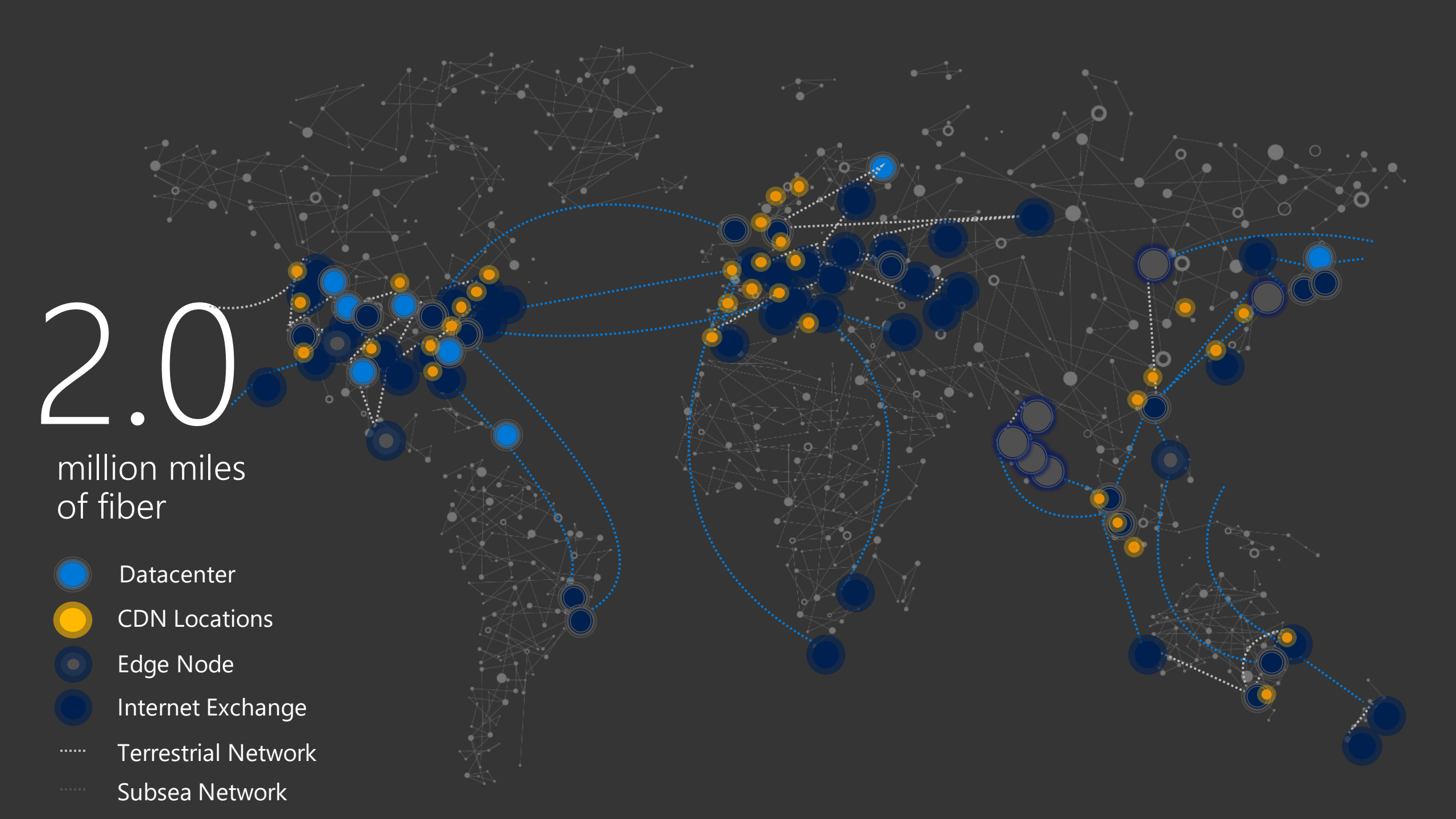


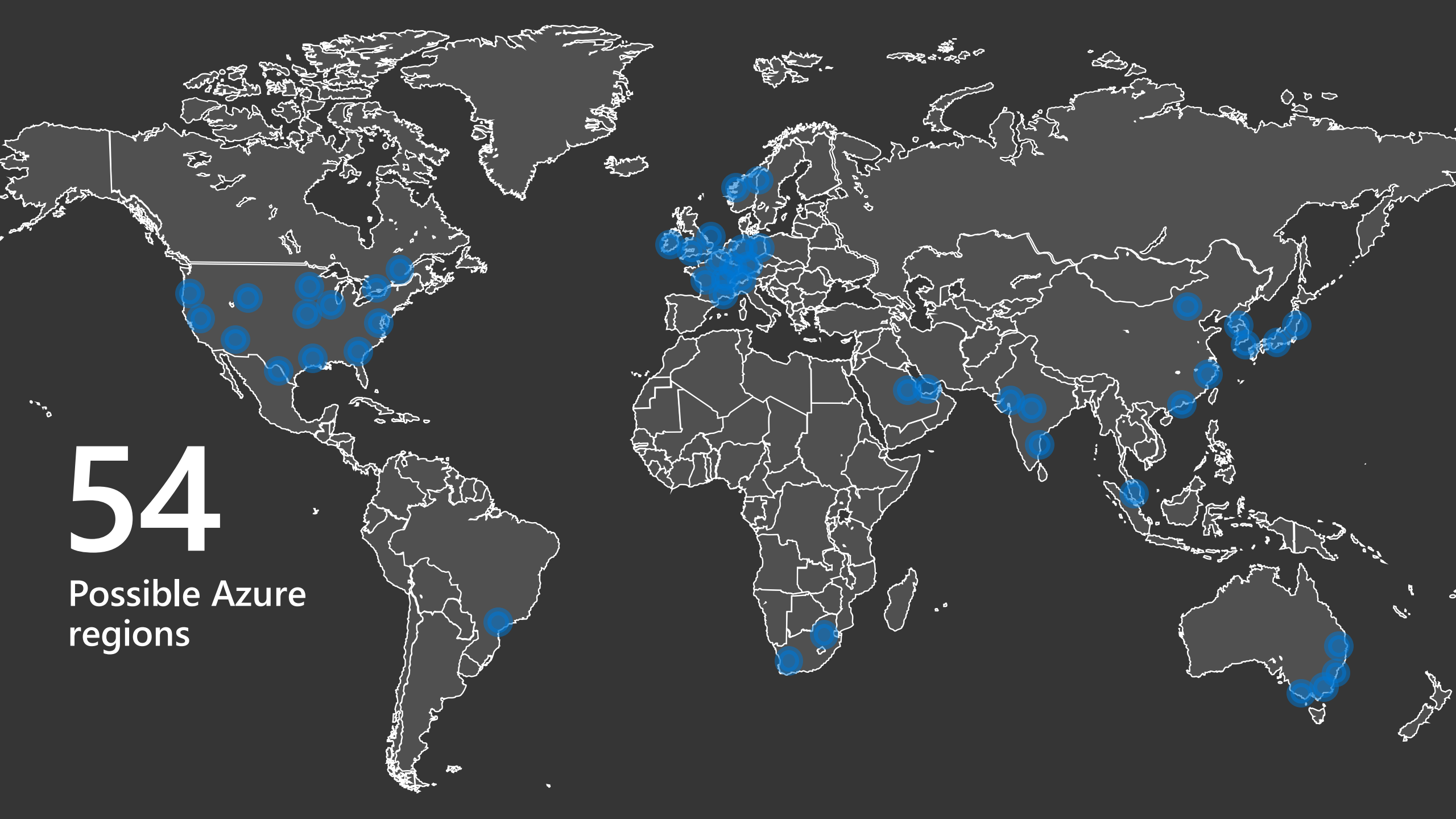
= 250,000 Libraries of Congress



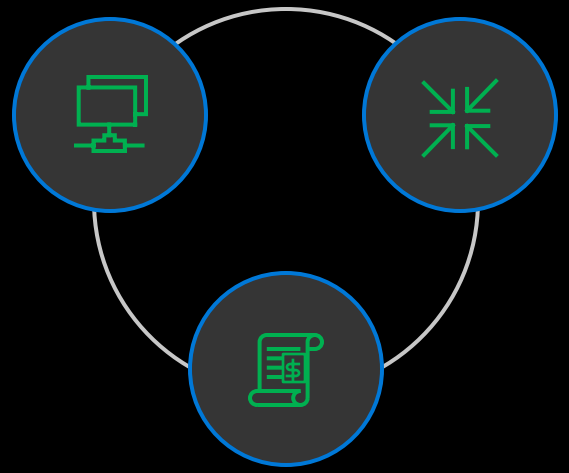


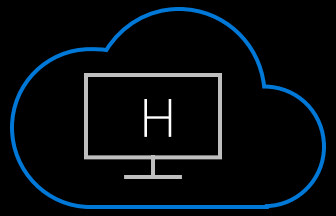
Analytics
Cryptography
Big Data
Genomics
Database
Network
CFD
Finance
FSI
FPGA
EDA
Visualization
Storage
Compute
System Validation
Bioinformatics
Healthcare
AI
Productivity
Scale
Simulation
Database
Media
Defect Inspection
XILINX
VIRTEX
UltraScale+
XCVU13P
FHGB2104AAZ


2.0
million miles
of fiber
Datacenter
CDN Locations
Edge Node
Internet Exchange
Terrestrial Network
Subsea Network


54
Possible Azure regions

Azure Specialized Infrastructure





High-performance VMs

Tightly coupled parallel jobs with InfiniBand

Accelerator-enabled VMs

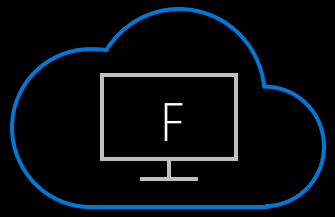
NV—Graphic-based applications
NC—GPU Accelerated Compute
ND—Deep Learning
NP—General Purpose FPGA

Cray in Azure

Managed, Custom Bare-metal
HPC or Supercomputing
On the Azure Network

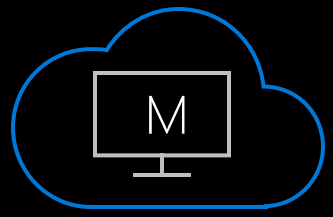
>80,000 IOPs Premium Storage

Low latency, high throughput apps



Compute-optimized VMs

Batch processing, Monte Carlo simulations



Large memory VMs

Large databases

Azure FPGA Runtime Platform – NP series

- Develop today, deploy tomorrow

- Write anywhere, run everywhere:
  - Standards-based platform:
    - Xilinx Alveo U250 board platform
    - Vitis shell and runtime (2020.2 XDMA (2.1))
    - RTL, HLS (C++), or mixed-language
    - DMA I/O to host via PCIe Gen. 3 x16

- Best-in-class PR shell, 80%+ device available to user
- 64 GB of DDR4 2400 (4 banks) per accelerator

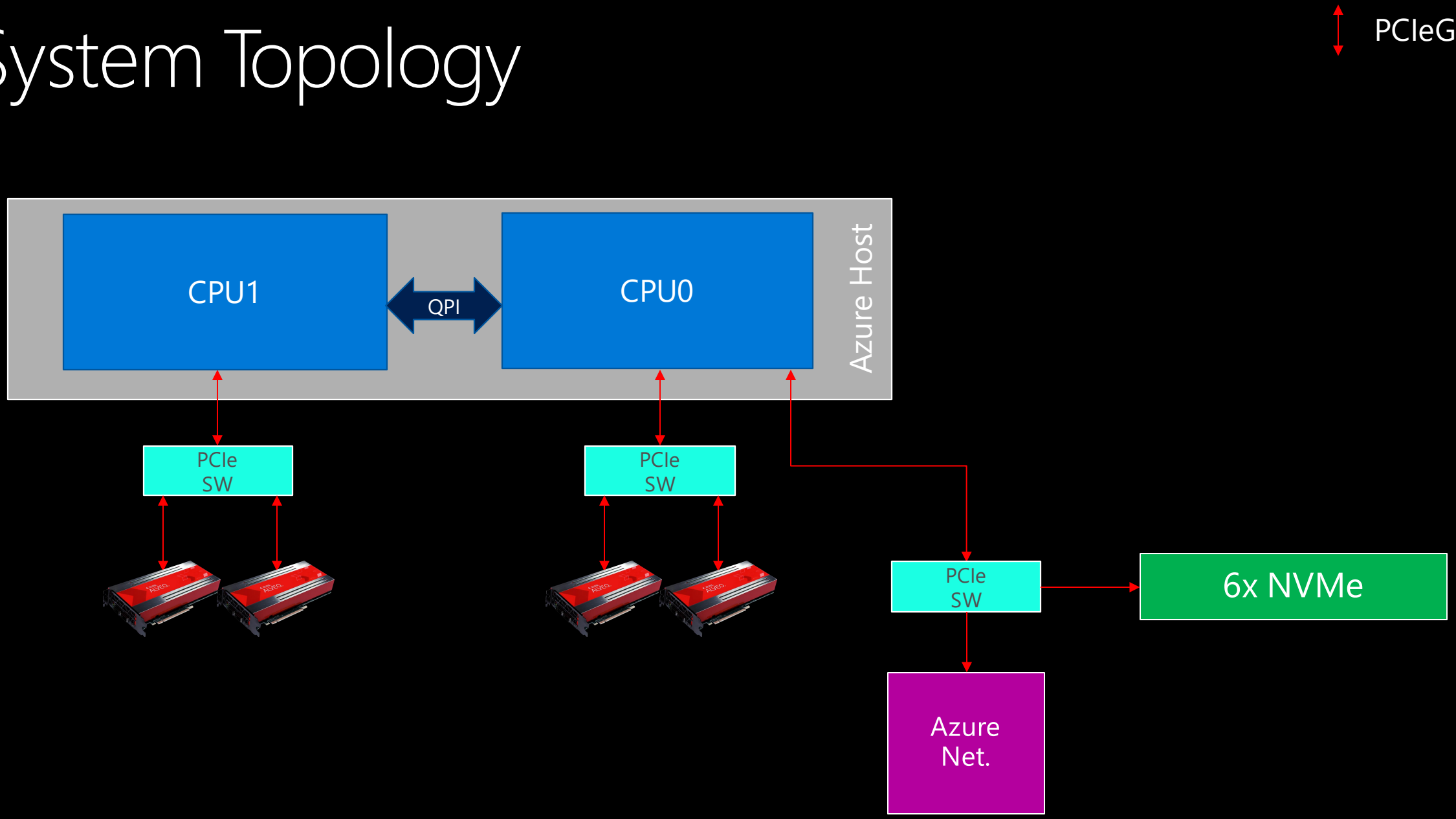
Dramatic Perf / TCO benefits in:
Genomics & Life Sciences
Computational Chemistry
Physical Simulation
Transcoding & Postprocessing
Analytics & Databases
Monte Carlo & Risk Analysis
Systems Behavior Validation
Computer Vision
Hardware Simulation and EDA

VM Specifications	NP10	NP20	NP40
Physical Cores (Intel Skylake)	10	20	40
Temp. High-Speed Storage	0.7 TB	1.4 TB	2.8 TB
Host RAM	168 GB	336 GB	672 GB
Accelerators (U250s)	1	2	4

Maximum System Specifications

- Two 26-core Xeon Platinum (Skylake) CPUs 2.6 GHz – 3.3 GHz
  - Up to 40 cores fully available to users
  - Configured for demanding workloads:
    - Turbo-Boost
    - No deep Cstates
    - No HyperThreading
- 672 GB DDR4 Memory available to users
- 50 Gb Ethernet w/ Azure second Gen SmartNIC
- Azure Standard & Premium Storage w/ Temp Disk backed by 6x NVMe SSDs
- 4x Alveo U250 Accelerators with 64 GB DDR4 each
- Vitis 2020.2 on RedHat, CentOS or Ubuntu

System Topology



Disk storage on NP Series VMs

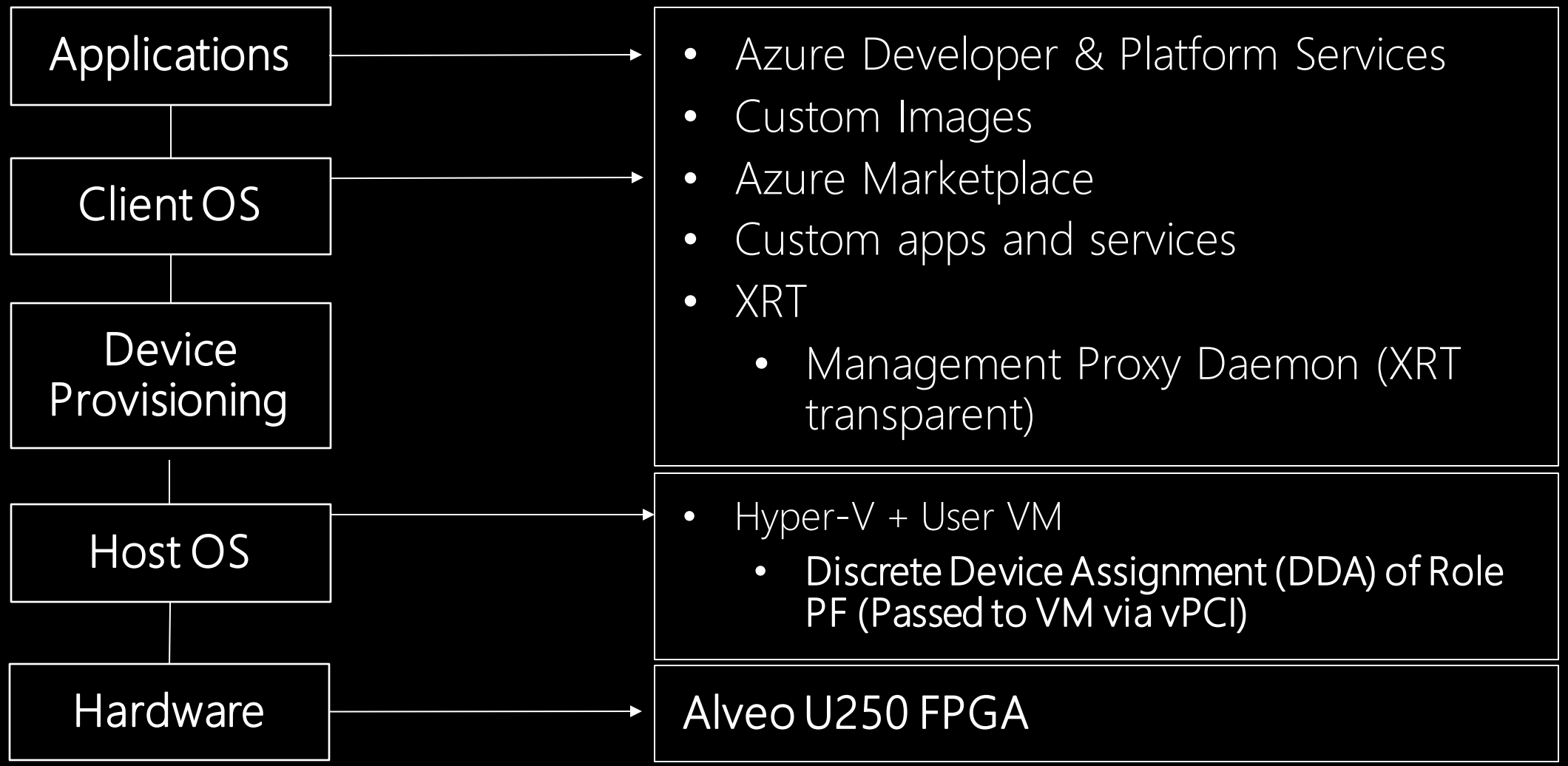
6x NVMe SSDs “Under the hood” dedicated to your data.

1 TB NVMe SSD	1 TB NVMe SSD	1 TB NVMe SSD
1 TB NVMe SSD	1 TB NVMe SSD	1 TB NVMe SSD

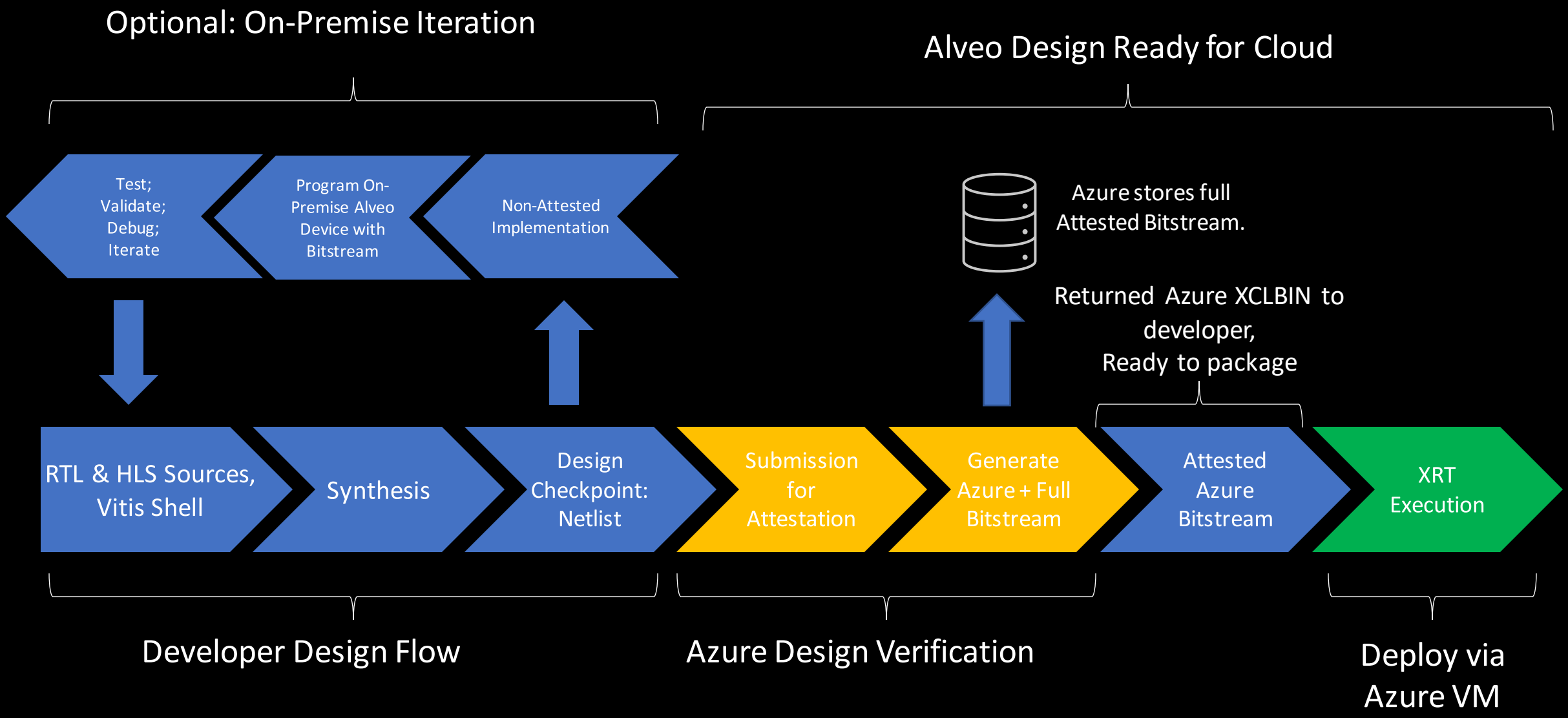
0.7 TB to 2.8 TB exposed to your VM via AHCI

0.7 TB to 2.8 TB used as cache for Premium SSD

Virtualizing the Accelerator: Under The Covers



Design Flow – Building Azure-Ready Alveo Solutions



Attestation Service

Upload netlist to Azure blob storage

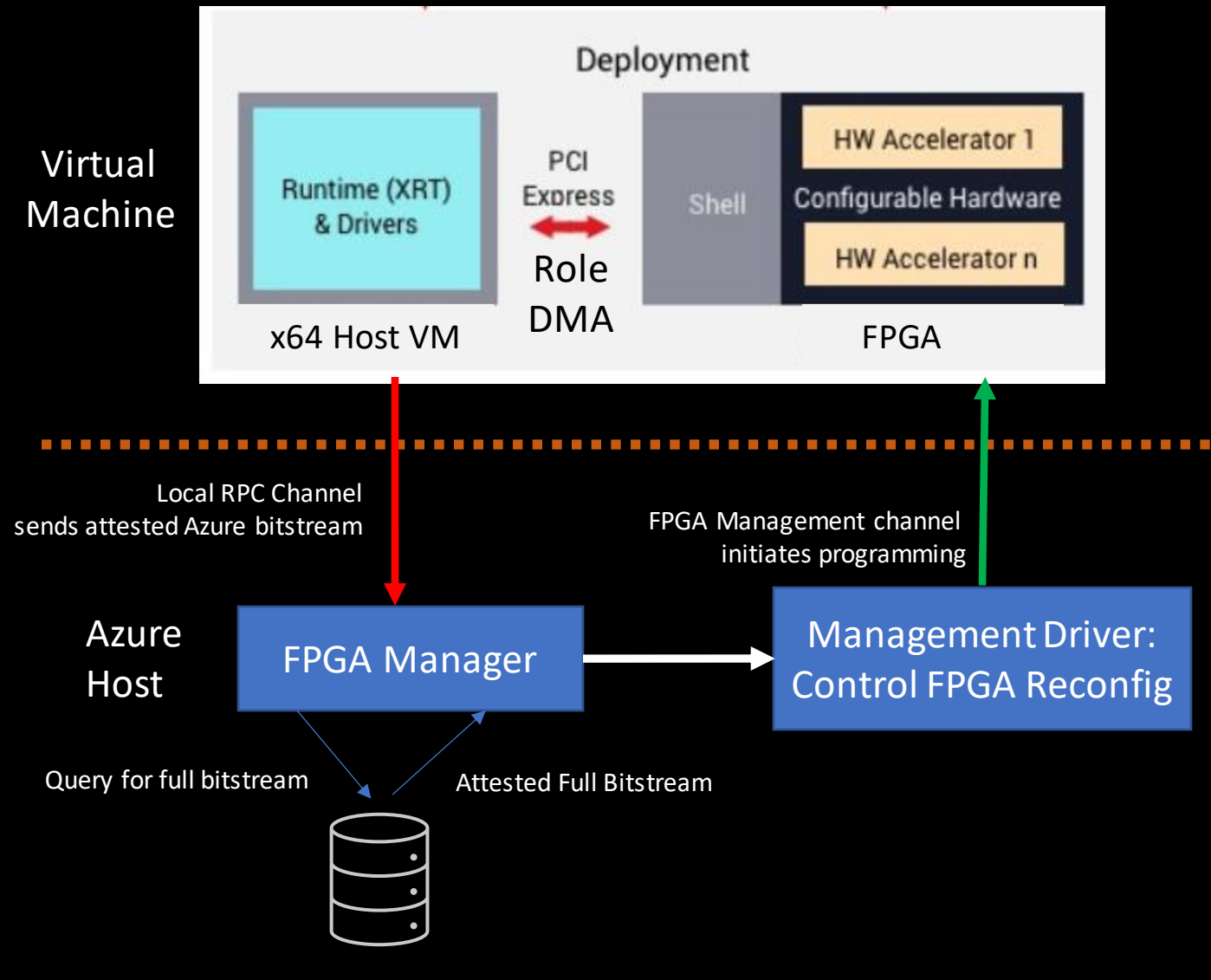
Run provided script to submit design to Attestation Service

Check shell compatibility and platform security

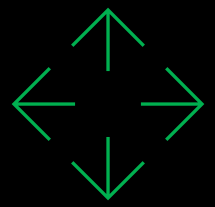
Abstract bitstream for IP protection and allow for safe sharing

Under the Hood – Deployment with XRT

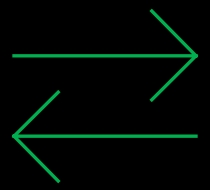
Reconfiguration control flow orchestrated by trusted Azure Host



Cloud Value for FPGA Workloads



Scale



Elasticity

Pay for use

Reach & locality

Hardware selection

No infrastructure

Known costs

The new Accelerated Opportunity

Build your Accelerated solution in the Cloud

Remember…

- Today's Vitis 2020.2 Alveo U250 solutions are ***tomorrow's Azure solutions:***
  - Lift and Shift
  - One design, on-prem or in the Cloud

Remember…

- Today's Vitis 2020.2 Alveo U250 solutions are ***tomorrow's Azure solutions:***
  - Lift and Shift
  - One design, on-prem or in the Cloud

- ***Tomorrow's Azure solutions*** get the benefits of the Cloud *without* additional engineering effort.

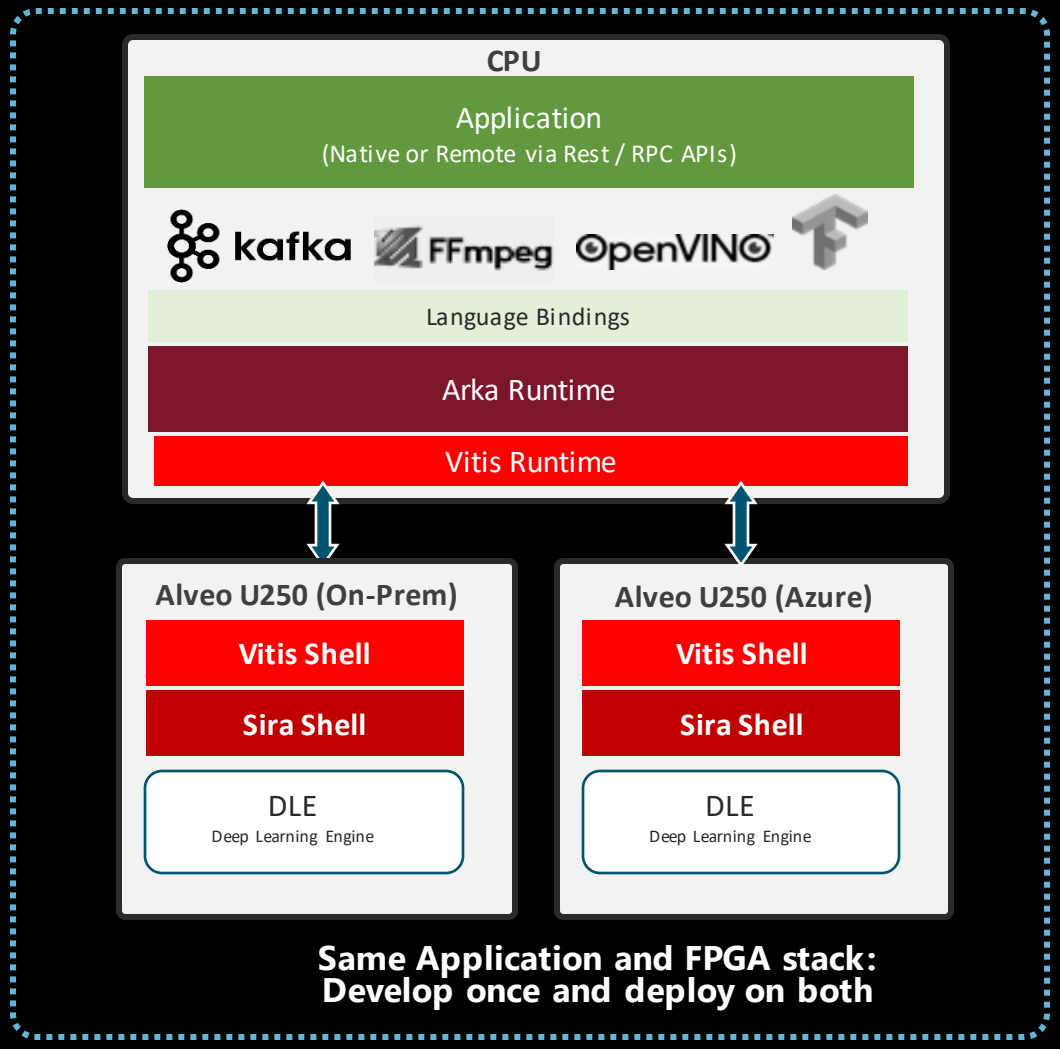
Remember…

- Today's Vitis 2020.2 Alveo U250 solutions are ***tomorrow's Azure solutions:***
  - Lift and Shift
  - One design, on-prem or in the Cloud

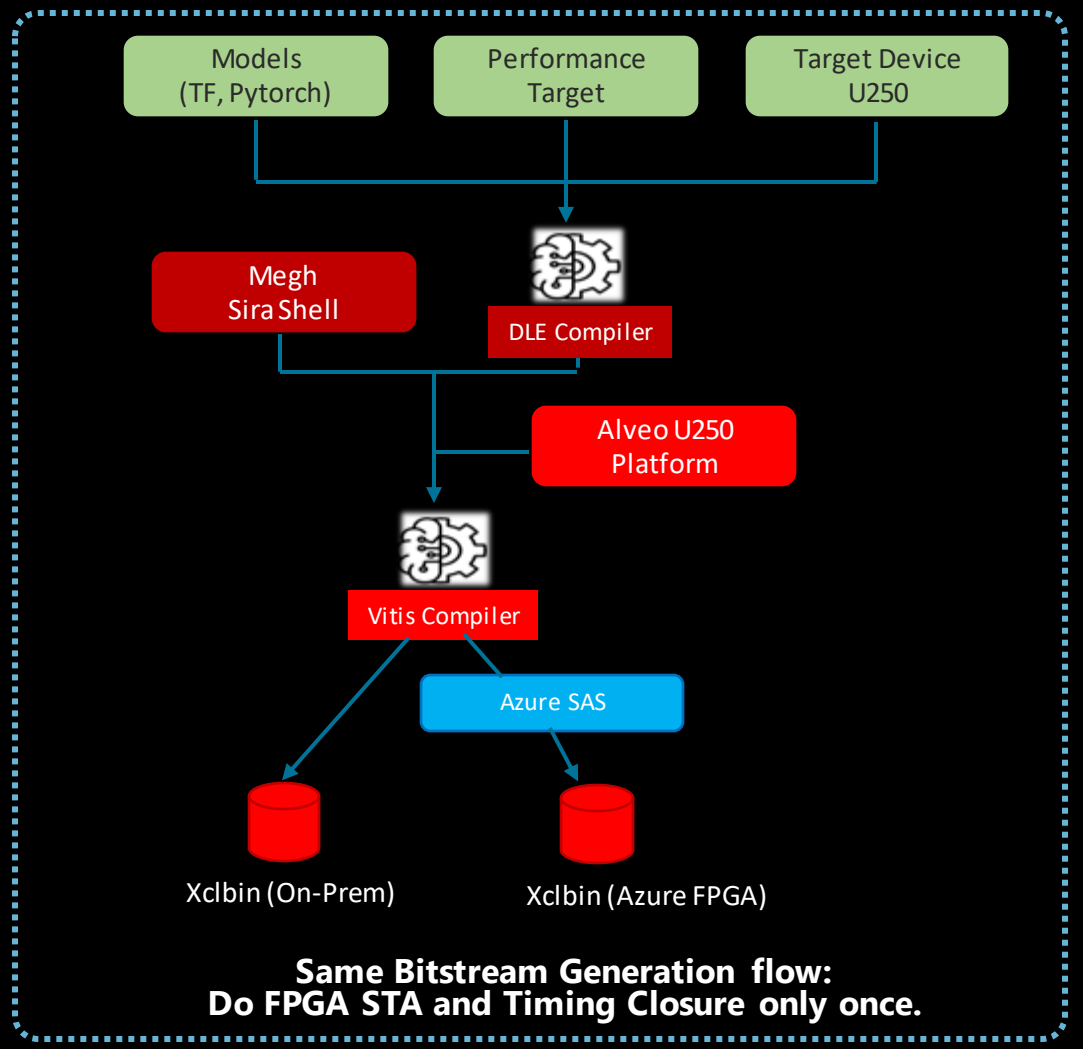
- ***Tomorrow's Azure solutions*** get the benefits of the Cloud *without* additional engineering effort.

And today, we're excited to showcase how…

Transitioning Megh DLE (Deep Learning Engine) from On-Prem to Cloud



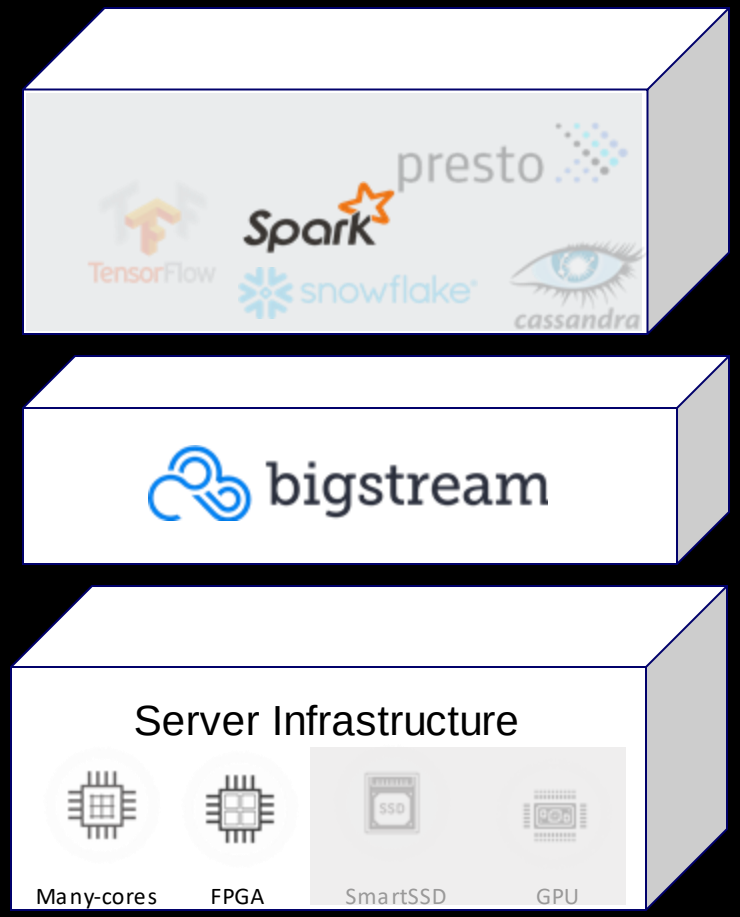
Same Application and FPGA stack:
Develop once and deploy on both



Same Bitstream Generation flow:
Do FPGA STA and Timing Closure only once.

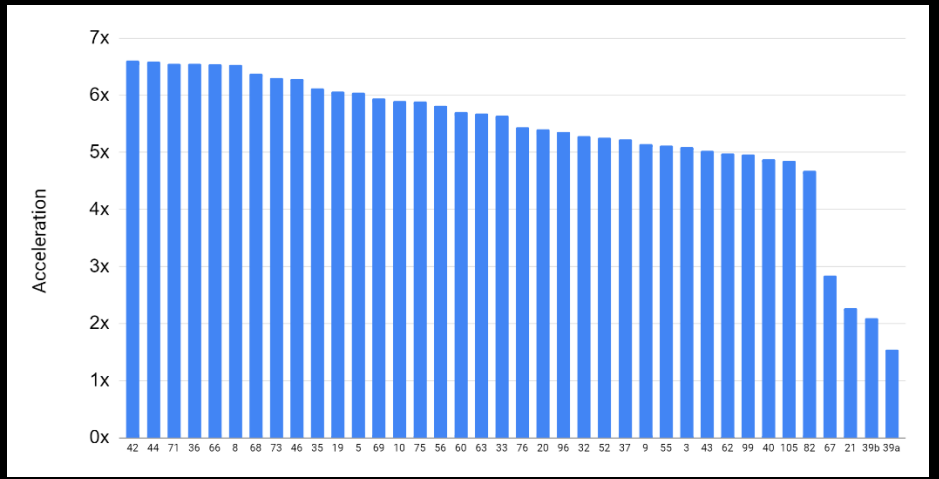
Common Vitis Flow across On-premise and Azure Cloud enables zero friction transition

Bigstream big data acceleration: Azure + Alveo



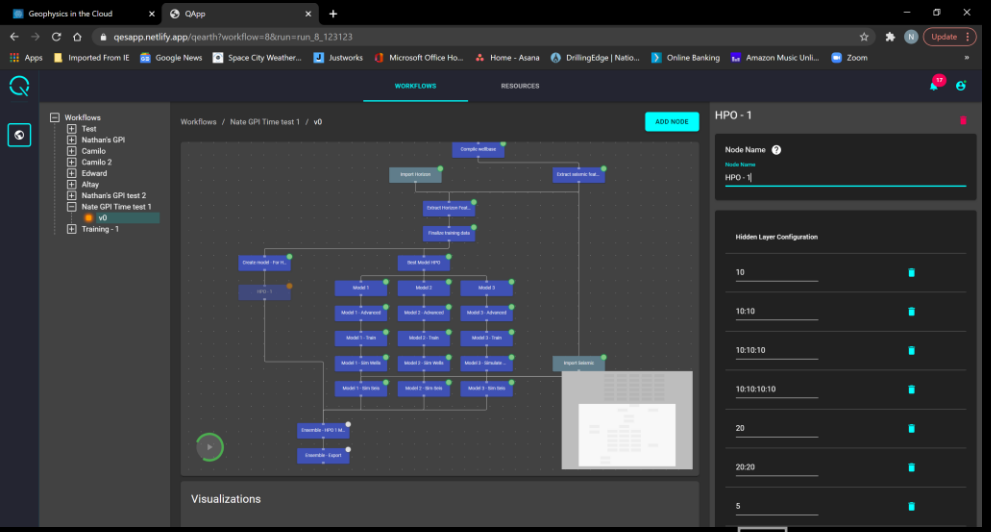
- Key use cases: ETL, batch analytics
- Bigstream: automates programming of Spark onto accelerators such as FPGA
- On-prem + cloud
  - **Azure NP10, NP20, NP40 (Alveo U250)**
  - **Streamlined migration**

Initial Results - 38 TPC-DS Queries
5.3x average acceleration (NP10)

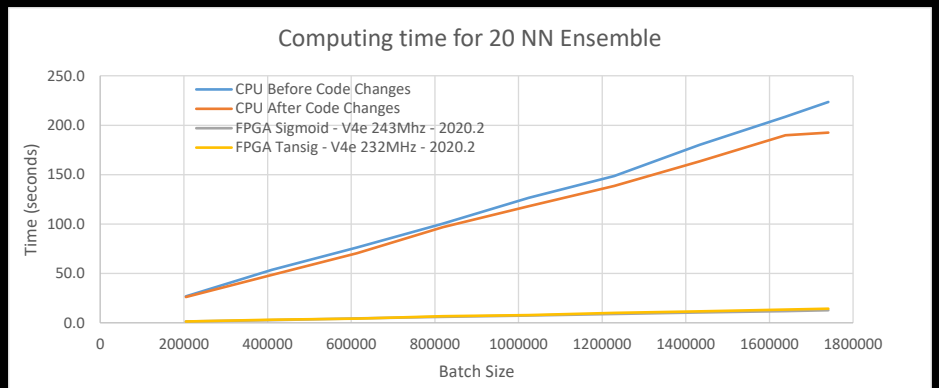


bigstream

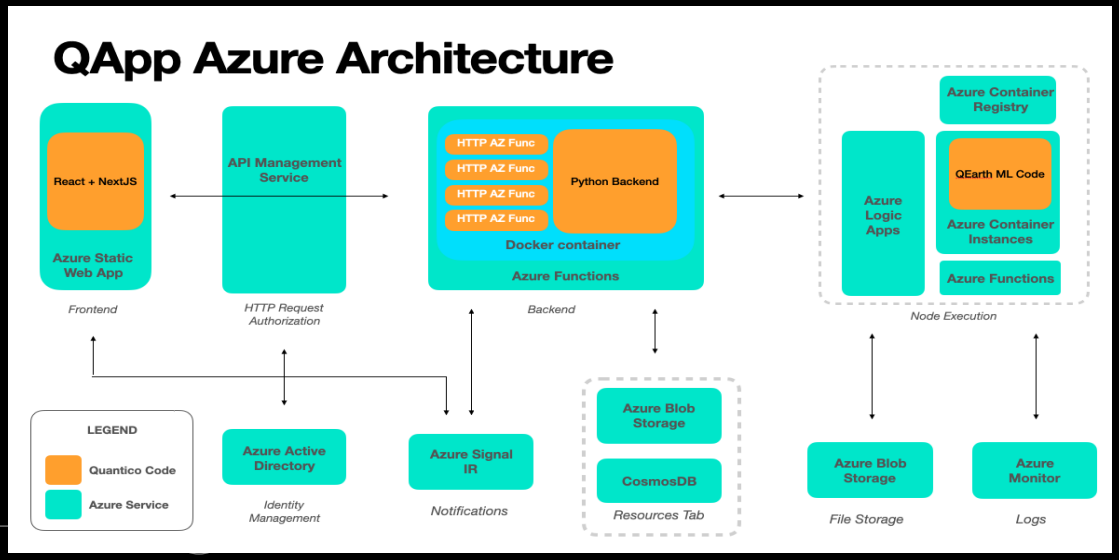
Cloud Native earth modeling application



On premise hardware acceleration – 16X faster prediction workloads



	Batch size			
20 NN	1433600	1638400	1740800	Average
Tansig - CPU (seconds)	181.0	209.6	224.2	
	163.6	188.6	191.9	
Xilinx - Tansig (seconds)	11.7	13.5	14.4	
Xilinx - Sigmoid (seconds)	10.5	12.0	12.7	
Xilinx - CPU - Tansig Acceleration	14.04	14.01	13.30	14.80
Xilinx - CPU- Sigmoid Acceleration	15.56	15.86	15.15	16.35

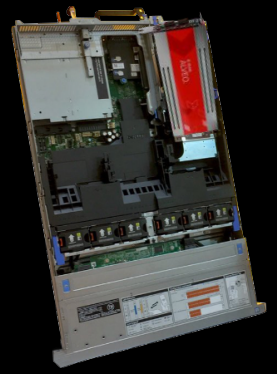


Microsoft Azure

NP-Family

DRAGEN Migration from Onsite to NP VM

Local Processing



Single Genome Analysis in 33 Minutes

(8 core machine with 1 FPGA)

- Bitstream Generation accomplished through Azure Attestation flow.
- Leveraging the same basic software and XRT communication
- Stream data straight from Azure Blob Storage.
- Use Shared Image Gallery to store VM with all software pre-installed
- Same Speed as DRAGEN onsite Server with same number of cores
- Porting from onsite to Azure was a smooth and easy process completed in approximately one month.

Microsoft Azure

Single Genome Analysis in 32 Minutes

(8 core machine with 1 FPGA)

illumina®

*Empirical analysis generated from >30 diverse genomes ; 1% ~ 3M bp, DRAGEN 3.7

© 2021 Illumina, Inc. All rights reserved.

Azure Synapse Analytics

The first unified, cloud native platform for converged analytics

Azure Synapse is the only unified platform for analytics, blending big data, data warehousing, and data integration into a **single cloud native service** for end-to-end analytics at cloud scale.

Azure Synapse Analytics

FPGA Accelerated Apache Spark

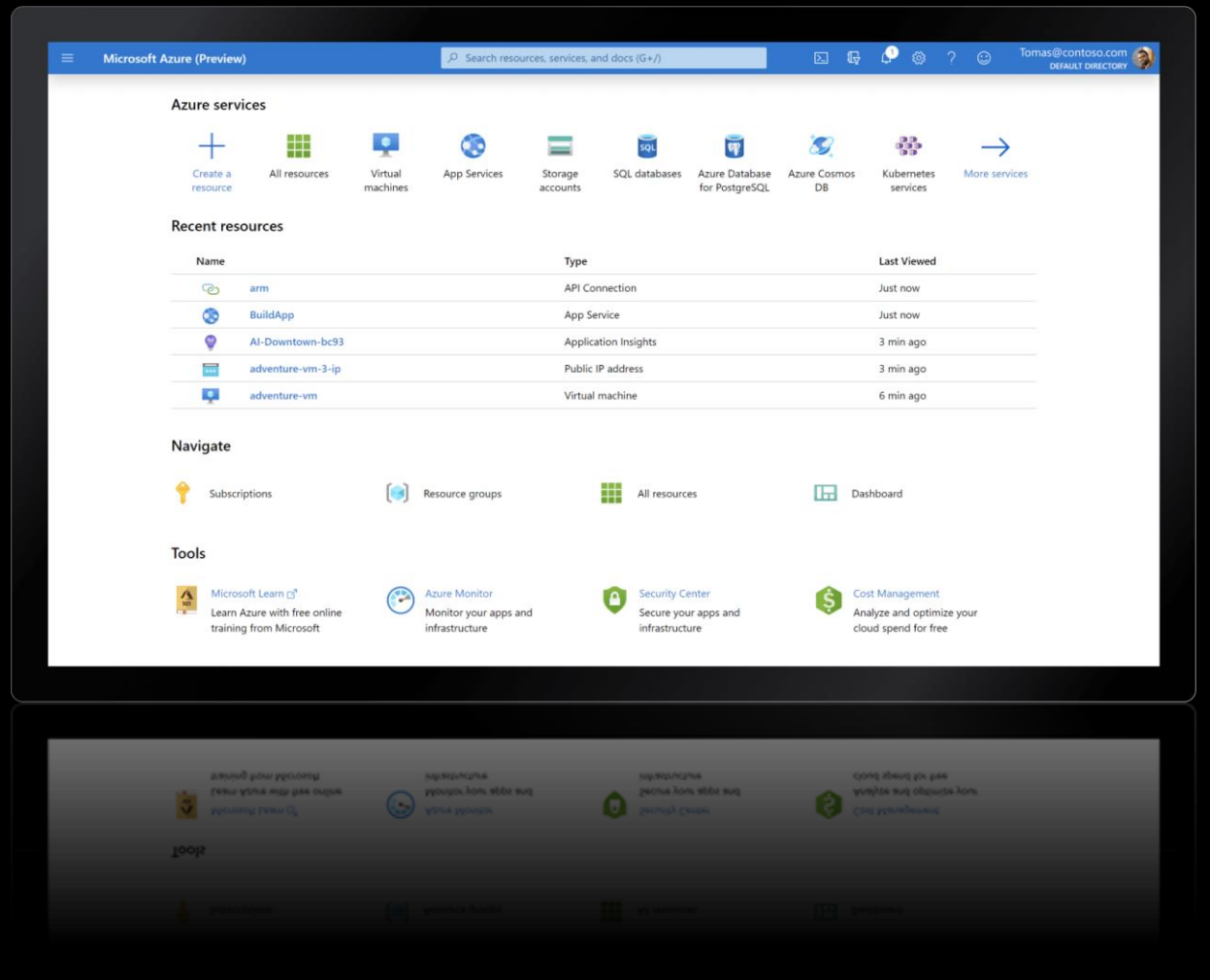
- Data volumes are continuously increasing
- Parsing takes up a significant part of the data prep workload in Apache Spark
- Collaboration with Azure NP-series team to develop FPGA accelerated CSV/JSON parser
  - Apache spark performance is 15MB/sec per physical core (Ev3 series)
  - FPGA CSV parser raw performance is 6.1 - 7.7 GB/sec. FPGA utilization of Parsing logic is 25% of U250.

Get ready to distribute your virtual appliances on NP-series VMs via Azure Marketplace.

NP VM Appliances are GA on 4/1

- Not required to use NP VMs.
- Simple, flexible, turnkey license terms and publishing model supports Pay-Per-Hour, and Bring-Your-Own-License.
- No different than publishing marketplace images on any other Azure VM Product.
- Same image can support non-Accelerated and Accelerated VMs.

Grow your reach with simplified deployment on Azure.



Availability & Pricing

NP Series VMs are GA in key Azure regions on 4/1.

Availability includes Pay-As-You-Go, as well as spot-pricing and reserved instances in:

- East US
- West US 2
- Southeast Asia
- West Europe

Anticipated rates for US, Pay-As-You-Go, On-Demand deployments beginning 5/1:

Subject to future change, availability, and regional premiums.

U250 Accelerators	Physical vCPU cores	Price Per Hour
1	10	$1.65
2	20	$3.30
4	40	$6.60

We can't wait to see what we'll build together.

Q&A

Get in touch:
AzureNPFeedback@service.microsoft.com